Outline for Uncertainty Analysis Workshop September 13, 2004
South Florida Water Management District
3301 Gun Club Road, West Palm Beach, Florida

Dr. Christine A. Shoemaker
Joseph P. Ripley Professor of Engineering
210 Hollister Hall
Cornell University
Ithaca, N.Y. 14853
607-255-9233(tel.); -9004(fax); -3328 (assistant) CAS12@cornell.edu

We propose to assess uncertainty in models of Everglades by a nested series of analysis procedures outlined below .

BACKGROUND: Let ***z*** be a vector of input values, ***x*** be a vector of model output, ***p*** be a set of parameters and ***H(p)z*** be the model being evaluated, so

$$\boldsymbol{x=H(p)z.} \qquad (1)$$

Let $\boldsymbol{P}=(P_1,\ldots,P_{D1})$ be the vector of the current values of the parameters that are assumed to be the best values (based on prior measurement or other analyzes). The symbol $\boldsymbol{p} = (p_1,\ldots,p_{D1})$ is used to denote a vector of variables that can take on different values. Boldface ***P ,p, z, x*** are used to denote vectors and boldface ***H(p)*** denotes a matrix that has elements dependent on ***p*** and other factors.

In all the analyses below, the computational effort is primarily for model simulations. The other computations for response surface fits, optimization and statistical analysis take seconds or at most a few minutes of computational time and hence are not significant in comparison to CPU time requirements for the simulations of ***H(p)z.*** The analysis system we have developed is **an integrated procedure for calibration, sensitivity analysis and uncertainty analysis** so that computationally expensive simulations done for one purpose can be reused for another analysis. (For example, I discuss below how simulations done for Part I are re-used for Part II.)

We are experienced in response surfaces (e.g. Chen et al., 1999) and in applying the integrated procedure to large complex models where the function is treated as a black box (e.g., we are given only the executable code) as in Regis and Shoemaker (2004 a and b in press). We have two currently funded NSF projects on use of response surface methods in calibration, sensitivity analysis, and uncertainty analysis and on parallel processing and grid computing. We also have access to the Cornell Theory Center, which is a high performance facility to do parallel computation when needed. We also expect to receive a third NSF proposal (done jointly with statistician Ruppert) on integrating response surface methods with Bayesian analysis for parameter estimation, sensitivity, and uncertainty analysis.

Part I: Identifying a small set of sensitive parameters:

Complex models have hundreds or even thousands of parameters that could be uncertain. Our first procedure is to examine the effect of changes in a large number of parameters on many model outputs. $\boldsymbol{D_1}$ is this number of parameters and it can be large. Our goal is to identify a smaller subset of parameters to which the model output is most sensitive and to analyze this smaller subset more thoroughly in Parts II and III. . Part I can be omitted if a small number of parameters has been selected upon which you wish to base the uncertainty analysis.

For the analysis in Part I, we consider the impact of changes in a single component P_k of $\boldsymbol{P}$ at a time. This approach uses a method described in Benaman 2002 and in manuscript Benaman and Shoemaker 2004 (paper attached) For each $P_{k,}$ we obtain from the literature, the allowable range of the parameter, which is based on the maximum (P_k^{max}) and minimum (P_k^{min}) allowable for each parameter (e.g. these are the upper and lower bounds on P_k).

Perturbation method: Sensitivity is based on changes in the values of model output in response to a perturbation in parameter. Perturbation of p_k can be based on one or both of the following methods: a) perturb p_k up and down by a fixed percentage or b) perturb p_k up where the perturbation is = frac * (P_k^{max}-P_k) where $0 \leq$ frac ≤ 0.5 and frac is a constant chosen by the analyst. Hence $P_k^+ = P_k +$ frac * (P_k^{max}-P_k). We then use a similar expression involving P_i^{min} to select the negative perturbation. The advantage of method b) is that it takes into account the uncertainty in the value of p_k (e.g. large value of (P_k^{max}-P_k) leads to a larger perturbation). The advantage of method a) is that it does not require the values of P_k^{max} or P_k^{min}, which are often not well known.

The number of simulations required for this analysis will be $2\boldsymbol{D_1}$ if just perturbation a) or b) is used and $4\boldsymbol{D_1}$ if both methods a) and b) are used. To simplify the following description, we will assume only one of the two perturbations methods is used since it is straightforward to see how to modify the description to deal with use of both methods.

We then perform all the perturbation simulations. Let P_k^+ and P_k^- be the values when P_k is perturbed by method a) or b) above. So we compute and store (P_k^+, $\boldsymbol{H}(P_k^+)\boldsymbol{z}$) and ($P_k^-$, $\boldsymbol{H}(P_k^-)\boldsymbol{z}$), where $\mathbf{H}(P_k^+)\boldsymbol{z}= \boldsymbol{H(p)z}$. where $\boldsymbol{p}$ is the base case except that the k^{th} component has been varied.

Recall that $\boldsymbol{H(p)z}$ is a vector-valued function that includes a number of model outputs $O_w(\boldsymbol{p})$ about which we want sensitivity and uncertainty analysis for w=1,…,Nk. For example $O_w(\boldsymbol{p})$ could be the water level at a specific location and time period for w=1 and $O_w(\boldsymbol{p})$ for w=2 could be water level at a different location or represent a different constituent like phosphorous.

The PDTs (Project Delivery Team(s) of CERP Comprehensive Everglades Restoration Project) can select the output variables $O_w(\boldsymbol{p})$ for which they want the uncertainty analysis reported. One simulation of the model H generates hundreds or thousands of outputs, so one needs to decide which Nk output he wishes to examine. It should be noted that with our approach, the computational effort is related primarily to the $\boldsymbol{D_1}$, and

increasing the value of Nk adds very little extra computation. Hence, with this method, it is possible to explore the uncertainty associated with a large number of different model outputs (i.e. a large Nk).

The next step in the Benaman and Shoemaker approach is to compute the individual sensitivity of the model for each parameter value P_k, k=1,…, $\boldsymbol{D_1}$. This sensitivity will depend upon the output variable(s) $O_w(\boldsymbol{p})$ selected. Hence to do the analysis, we can examine the sensitivity of changes in parameter P_k by looking at a sensitivity index. One sensitivity index we can use is $SI_{k,w}$ where

$$SI_{k,\omega} = \max\left\{\left|\frac{\frac{O_w - O_{kw}^{+}}{O_w}}{\frac{P_k - P_k^{+}}{P_k}}\right|, \left|\frac{\frac{O_w - O_{kw}^{-}}{O_w}}{\frac{P_k - P_k^{-}}{P_k}}\right|\right\} \qquad (2)$$

O_w^{+} and O_w^{-} are the values of the output function for P_k^{+} and P_k^{-}, respectively. So $SI_{k,w}$ is the rate of change in the Output $O_w(\boldsymbol{p})$ maximized over the positive and negative perturbations. O_w is the base case output $O_w(\boldsymbol{P})$. The purpose of including the O_w and P_k in the denominator is to normalize so that the differences in $SI_{k,w}$ for different parameters can be compared.

Cumulative Sensitivity Analysis: However, $SI_{k,w}$ in (2) only indicates the sensitivity with regard to one output $O_w(\boldsymbol{p})$. In assessing uncertainty, we typically are concerned about many different model outputs. To have an index of the impact on multiple outputs, we define the *cumulative sensitivity index* $cumSI_k^m$ to be

$$cumSI_k^m = \sum_{for\ all\ \omega} \beta_w^m\ SI_{k,\omega} \qquad (3)$$

where β_w^m is a weighting that adjusts for the relative importance of output variable O_w. For example, one weighting (m=1) might be related to the quality of the fit of the data at different locations and another weighting (m=2) might be related to prediction of events during certain seasons. If we want to consider just one output, then the weighting scheme can be structures so that $\beta_w^m = 1$ for one w and equals 0 for all other w. Hence, we can have different overall indices $cumSI_k^m$ for different weightings (m).

Ranking: The next step is to rank the value of $cumSI_k^m$ for each m. This ranking is shown in Table 4 in from Benaman and Shoemaker (2004)) for 4 different weightings of different outputs. As shown in Table 4, some of the parameters are shown to be important for many different weighting. These are the parameters that should be selected for the focus of the uncertainty analysis in Part I and II, which requires more computation per parameter considered than the individual sensitivity analysis discussed in Part I.

The further analysis will be based on *d* parameters, where d < D_1. The users will select how large d is. The computational difficulty of Part II and Part III will depend primarily on the magnitude of d. The analysis in Part I helps determine which are the most important parameters to include in d. Also the magnitude of the $cumSI_k^m$ gives a quantified value for the importance of each parameter for a given weighting and this enables the user to pick a d that is not unnecessarily large and thereby reduces computational effort associated with Parts II and III

Summary of Part I: This method has done a computationally efficient analysis of the individual sensitivity of a relatively large number of parameters in order to identify which *d* of these are most important for additional analysis in Part II and III. Identification of *d* parameters is also critical in model calibration. All the simulations involving perturbations of the d parameters will be re-used in the analysis in Part II and III. The total number of simulations done in Part I is $2D_1$. 2d of the simulations done in Part I can re-used in the analysis in Part II.

Part I can be omitted if PDT has already selected a small number of parameters upon which they wish to base the uncertainty analysis. However, if earlier sensitivity analysis has been done to select the most important d parameters (using our method or some other method), it is important that these earlier simulation results be saved so that they can be added to the function evaluations used in constructing the response surfaces below for Part II and III.

Part II: Construction of a Response Surface using Symmetric Latin Hypercube Design and Previously Evaluated points.

Assume that because of new measurements, we want to assess the variability associated with the value of several parameters P_k, k=1, d. Assume that we know the probability distribution of the values of P_k.

Our next step is to use a Symmetric Latin Hypercube Design (SLHD) to select N more points at which to evaluate ***H(p)z*** in order to construct a response surface so we will have the values (***p***, ***H(p)z***) at many points. We use a symmetric Latin Hypercube Design as the initial evaluation points since the symmetry condition improves the space-filling properties of a Latin hypercube (Ye et al. 2000).

Typically, we use d(d+1)/2 points for a response surface, where d is the dimension . Since we already have 2 d points from part 1, we just evaluate d(d+1)/2 -2d new points. So now we have evaluated ***H(p)z*** for d(d+1)/2 different values of the vector $\boldsymbol{p}=(p_1, \ldots, p_d)$.

We will then fit a response surface ***RH(p)z*** to these points. The purpose of the response surface is to provide an approximation of the function H that can be quickly evaluated. We can afford computationally to do hundreds of thousands of evaluations of the response surface whereas we can only do a few (hundred or hundreds of) evaluations of

the true function H. $\boldsymbol{RH_i(p)z}$ is an approximation of the i^{th} output of H(z) so $RH_1(z)$ is an approximation of the ERROR function and $\boldsymbol{RH_i(p)z}$ for i>1 is an approximation of the output functions $O_w(z.)$

We will use a radial basis function response surface (e.g. Ruhmann, 2003) with which we have had considerable experience. We will then do N additional function evaluations. This will be done using our response surface approaches using radial basis function (Regis and Shoemaker, 2004, in press). Kriging is one type of radial basis functions that can be used for a response surface. We will explore which type of response surface method is most effective. We will construct the radial basis function for all the output variables deemed to be important. For example the hydraulic heads of groundwater in a number of locations are all different output variables based on $\boldsymbol{H(p)z}$ and a response surface can be created for each of these outputs.

We will then estimate the mean H_i^{avg} and the variance H_i^{var} of **each** output variable $\boldsymbol{H_i(p)}$ using the following formulae:

)

$$H_i^{avg} = \sum_{j=1}^{NJ} prob(P^j) * RH_i(P^j) \qquad (3)$$

$$H_i^{\text{var}} = \sum_{j=1}^{NJ} prob_1(P^j) * [RH_i(P^j) - H_i^{avg}]^2$$

The term $prob\ (P^j)$ is the joint probability that the value of the vector is P^j which is known from the measurement values and an assumption of independence. We use the notation $H(P^j)$ to indicate the j^{th} combination of parameter values which is a specific instance of the vector $\boldsymbol{p}$. If only one parameter P_k is being varied, then P^j has the same components as the base case parameter vector $\boldsymbol{P}$ except that the k^{th} component has been perturbed.

In this case NJ is a large number (e.g. 100,000) and represents a grid partition of the d dimensional space of the parameter variables. The averages and the variances are then weighted by this probability. Because the response surface is so efficient to evaluate, the evaluation at 100,000 points is not computationally expensive in comparison to the evaluation of a single simulation.

Because we will have done hundreds of thousands of evaluations of the response surfaces, we can also compute the cumulative distribution function to describe the uncertainty (which is more information than just giving the variance) and determine the probability that a certain variable exceeds some critical threshold { (which threshold can be defined by PDT) }, which might be more important and practical than just knowing the variance.

We can also use this approach to estimate the error in the response surface. We can use this information to guide the selection of additional points for evaluation of the real function H(z). If NZ more points are selected to be evaluated, then the response surface would be updated and the estimates in (3) would be recomputed.

Part III: Bayesian Analysis: In this step we will use a more complex analysis involving Bayesian statistics that will estimate the probability distribution and covariances associated with each of the d uncertain parameters. In the process, the mean and variance is also computed. This is the topic of a new NSF project that has just been accepted (apparently) by NSF that is done in collaboration with a statistician David Ruppert. We would apply this analysis to the simulations done for Part II. The prior distribution can be either assumed to be uniformly distributed within its allowable range or the prior distribution can be based on the experimental data used to estimate the parameter. The method also incorporates transformations to deal with distributions that are not symmetric. The method also nests the analysis. In particular we propose to first do the analysis only for water movement (e.g. stages, velocities, etc.) and then use the posterior distribution of the water parameters as the prior distributions for parameters related to constituents like phosphorous. This can be done because the phosphorous parameters do not affect water movement, but the water parameters can affect chemical constituents like phosphorous.

Once we have the probability distributions of the parameters, we can then do Monte Carlo sampling on the response surface using the given probabilities to generate output for uncertainty and from this also assess the probability that a critical threshold is exceeded. This analysis does a more thorough analysis of the statistical properties of the parameters and such information adds to understanding.

It is expected that to get a reasonable parameter distributions that Part III may take more simulations that Part II, in which case we would need to increase N (from part II). We propose to do experiments with test functions (that are computationally fast) to estimate how much larger N would need to be for the Bayesian analysis. The PDT would need to decide if the extra information gained from these simulations is worthwhile. The extent to which the more extensive statistical analysis proposed in Part III should be implemented can be decided in further discussions with PDT.

Previous Experience of Investigator

The investigator has years of experience in water resource modeling and analysis related to optimization (e.g. Yoon and Shoemaker, 1999 and 2001; Mansfield and Shoemaker, 1998; response surfaces (e.g. Chen et al. 1999); calibration (Benaman et al., 2004), sensitivity (Benaman and Shoemaker, manuscript) and uncertainty analysis (e.g. Minsker and Shoemaker, 1998). This research has included applications to both groundwater models (including reactive transport models) and to watersheds.

Her research is currently funded by two NSF grants and a third grant has apparently been approved:

1."Improving Calibration, Sensitivity and Uncertainty Analysis of Data-Based Models of the Environment, "$350,000, 4/1/03 –3/30/06, [from Environmental Engineering program in Engineering. Directorate] **National Science Foundation**

In this project, new function approximation (serial) algorithms developed by my group will be applied to the problem of calibration of complex, nonlinear, computationally expensive, environmental models. As an example, the function approximation algorithm will be used with the transport model for anaerobic bioremediation of chlorinated ethenes by my group to assess the model's ability to fit field data. The methodology developed also permits the analysis of multivariate sensitivity and uncertainty analysis.

2. "Multi-Algorithm Parallel Optimization of Costly Functions", $380,000, 7/2003-6/2006 [from Advanced Computing Research Program in the Computer and Information Science Directorate] **National Science Foundation**.

The objective of this project is to develop an effective parallel algorithm for finding near optima for costly nonconvex black box functions f(x) for x in $D \subset R^d$ when derivatives of f(x) are unavailable. This methodology focuses on "costly" f(x), i.e. the CPU time to evaluate f(x) once can range from many minutes up to many hours or days. Such problems arise in many areas of science and engineering, including the optimization of nonlinear systems that are described by partial differential equations. Our approach is to design a coarse-grained procedure that is scalable and robust for a variety of application problems and computing environments. The proposed procedure iteratively uses function approximation algorithms. The algorithm will be applied to a range of difficult test problems and to costly real engineering functions. These applications come from the PI's own research projects on environmental pollution and safety of drinking water

3. "Integrating Bayesian Statistical Inference and Response Surface Optimization for Computationally Expensive Environmental Models", (with D. Ruppert) $660,000 1/1/05-1/1/08 [from the Statistics Program in Mathematical Sciences] {This proposal has been recommended for funding but we are awaiting final approval.]

The research results from these current NSF projects can be integrated into the analysis of the Everglades model analysis as appropriate. Hence, the Everglades project will benefit from newly developed research that is directly applicable to the analysis of computationally expensive models like those being used to describe the Everglades. The fact that these NSF projects have all been funded supports the high quality of the research in this area.

Shoemaker's research has been acknowledged by the receipt of a number of awards including election to grade of "Fellow" in the American Geophysical Union, receipt of the Hines Award from ASCE, receipt of a Humbolt Research Prize, and election to the Joseph P. Ripley Professorship in Engineering at Cornell University.

References

- Buhmann, M.D. (2003) *Radial Basis Functions*, Cambridge University Press, U.K.

- Koehler, J.R. and Wen, A.B. (1996) *Computer Experiments* in Ghosh, S. and Rao, C.R.., Handbook of Statistics, 13: Design and Analysis of Computer Experiments (DACE), North Holland, Amsterdam, pp. 261-308.

- Minsker, B.S. and C.A. Shoemaker, Quantifying the Effects of Uncertainty on Optimal Groundwater Bioremediation Policies, **Water Resources Research** 34, 3615-3625, 1998

- Mansfield, C.M and C.A. Shoemaker, "Use of Numerical Sparsity in an Aquifer Quality Optimization algorithm," **ASCE Jn. Of Water Res. Plan. Manag**. 124, 15-21, 1998

- Yoon, J-H and C.A. Shoemaker, "Comparison of Optimization Methods for Groundwater Bioremediation," **ASCE Jn. of Water Res. Plan. Manag.** 125,pp. 54-63, 1999.

- Chen, V. C. P., D. Ruppert, and C. A. Shoemaker, "Applying experimental design and regression splines to high-dimensional continuous-state stochastic dynamic programming". **Operations Research** 47 (1), 38-53, 1999.

- Yoon, J-H and C.A. Shoemaker, "Improved Real-Coded Genetic Algorithm for Groundwater Remediation", ASCE **Jn. of Computing**, July 2001.

- Benaman, J., and C. Shoemaker, "An Analysis of High-flow Sediment Event Data for Evaluating Model performance", **Hydrologic Processes** (accepted 2003)

- Benaman, J., C. A. Shoemaker and D. A. Haith, "Modeling Non-Point Source Pollution Using a Distributed Watershed Model for the Cannonsville Reservoir Basin, Delaware County, New York", **ASCE Jn. of Hydrologic Engineering** (in press 2004)

- Benaman, J. and C.A. Shoemaker, "Model Sensitivity Analysis for Large Numbers of Parameters," (manuscript, 2004)

- Regis, R. and C. Shoemaker, "Local Function Approximation in Evolutionary Algorithms for the Optimization of Costly Functions," **IEEE Trans. on Evol. Computation** (in press 2004a)

- Regis, Rommel, and C. Shoemaker, "Efficient Constrained Greedy Response Surface Methods for Global Optimization," **Journal of Global Optimization** (in press 2004b)

- Ye, K. Q., W. Li, and A. Sudjianto (2000) Algorithmic construction of orthogonal symmetric Latin hypercube designs, *Jn. Of Statistical Planning and Inference* 90: 145-159.